LLNL Environmental Restoration Division (ERD)
Standard Operating Procedure (SOP)

ERD SOP 5.20: Cost Effective Sampling (CES) Algorithm Preparation—Revision: 0

AUTHOR(S):
D. MacQueen

ENVIRONMENTAL RESTORATION DIVISION
CONTROLLED COPY
_____
THIS IS A RED STAMP
ERD

APPROVALS: **Date**

Albert L Lamarre 02/02/01
Division Leader

1/31/01
Environmental Chemistry and Biology Group Leader

CONCURRENCE: **Date**

Valerie Dibley 1/26/01
QA Implementation Coordinator

1. PURPOSE

The purpose of this SOP is to prepare statistically based ground water sampling frequency recommendations. The method used to produce the recommendations is referred to as the Cost Effective Sampling (CES) Algorithm.

The algorithm is based on the following principles applied to a database of previously measured chemical concentrations in ground water:

- The greater the measured trend (average change in concentration per year), the greater the sampling frequency.
- The greater the measured variability, the greater the sampling frequency.

2. APPLICABILITY

Applicability of the CES algorithm to particular wells and analytes is decided on a case by case basis by ERD project staff. In general, it is applicable to wells being used for routine monitoring

of known ground water contaminants, and is not applicable to wells in active use for remediation, such as pump and treat wells.

This document focuses primarily on tasks performed by the Algorithm Custodian (see Section 0) while performing an algorithm run (see Section 0), and to a lesser extent on tasks performed by the Quality Control (QC) Chemist.

3. REFERENCES

3.1 Dibley, V. R. and R. Depue (2000). SOP-2.11, "Developing Ground Water Monitoring Sampling Plans" (UCRL-109115 Rev. 8).

3.2 Ridley, M. R. (1999). Cost-Effective Sampling of Groundwater Monitoring Wells: A Data Review & Well Frequency Evaluation (UCRL-AR-133310).

4. DEFINITIONS

The following definitions are specific to this procedure:

CES Algorithm	The Cost Effective Sampling (CES) Algorithm is a system that is used to prepare statistically based ground water sampling frequency recommendations.
CES Algorithm run	The use of the CES Algorithm computer codes to produce sampling frequency recommendations.
CES Cutoffs	The statistical criteria used by the CES algorithm to determine the sampling frequency.
CES Final recommendation	A sampling frequency recommendation made by the project staff after consideration of the QC recommendation.
CES Planning quarter	The quarter for which a sampling plan is being prepared.
CES Previous quarter	The quarter before the planning quarter.
CES QC recommendation	An algorithm-based recommendation after review by the QC chemist.
CES Raw recommendation	A recommendation produced by the CES algorithm computer program, before review by any human being.
CES Working directory	The Unix directory containing the computer files for a particular CES algorithm run.

5. RESPONSIBILITIES

5.1 Algorithm Expert

The algorithm expert is the individual most responsible for developing and evaluating the algorithm.

5.2 Quality Control (QC) Chemist

For wells and chemicals where the algorithm recommendations are used as minimum-permitted sampling frequencies (see ERD SOP-2.11), only the QC chemist is authorized to reduce the sampling frequency relative to the raw recommendation. The QC chemist is responsible for identifying cases where the algorithm failed, and providing recommendations for those cases.

5.3 Statistician

The statistician works with the algorithm expert developing the algorithm.

5.4 Project Staff

The project staff are the scientists and engineers responsible for deciding which wells and chemicals are submitted to the CES algorithm, and for using the algorithm recommendations to produce a sampling plan.

5.5 Sampling Coordinator

The sampling coordinator is the individual responsible for scheduling the sample collection effort.

5.6 Sampling Plan Custodian

The sampling plan custodian maintains lists of algorithm wells and an electronic version of the final sampling plan in the database. See ERD SOP-2.11.

5.7 Algorithm Custodian

The algorithm custodian is responsible for performing routine algorithm runs. The algorithm custodian is responsible for computer source code development and maintenance, under the direction of the algorithm expert and the statistician. This procedure assumes that the Algorithm Custodian is familiar with the Unix operating system, SAS software, the EPD database, and SQL database programming.

6. PROCEDURE

The algorithm is run for the Livermore Site each calendar quarter (the “previous” quarter) to generate recommendations for the following calendar quarter (the “planning quarter”). The algorithm is run as needed at Site 300. Normally, an algorithm run is initiated about four to six weeks before the beginning of the planning quarter.

6.1 Procedure Overview

Steps to perform an algorithm run are these:

1. If necessary, the project staff review the list of wells to which the algorithm is applied and the sampling plan custodian updates the list.

2. The algorithm custodian runs computer programs to retrieve data and calculate raw recommendations.

3. The QC chemist reviews the raw recommendations. He or she identifies cases where the algorithm failed (if any) and makes recommendations for such cases.

4. The algorithm custodian prepares the recommendations for review by the project staff.

5. Project staff review the recommendations to identify wells at which a different sampling frequency is appropriate (see ERD SOP-2.11).

6. Project staff submit the final sampling decisions to the sampling coordinator and sampling plan custodian.

7. The algorithm custodian stores the raw recommendations in the database.

See Section 0 for a more detailed description of the steps performed by the Algorithm Custodian and QC Chemist.

6.2 Computer Resources

The software components of this system include:

- Analytical results from previous ground water samples, stored in a relational database management system (RDBMS).
- A current sampling plan stored in the database.
- Computer programs that perform the algorithm calculations.
- Data transfer capability between the algorithm software and the database.
- Spreadsheets.
- A World-Wide-Web (WWW) server.
- Scripts used by the WWW server to display the algorithm results.

At present the database is maintained using Ingres® software. The algorithm calculations use SAS® software, and the spreadsheets use Microsoft Excel.®

The computer hardware components of this system include:

- Unix workstations.
- Personal “desktop” computers.
- Hard disk space sufficient to store the necessary computer files.
- Network connections between the Unix and desktop systems.

At present the Ingres database and the SAS software run on Sun Microsystems SparcStation® Unix systems. The spreadsheets reside on Macintosh® computers.

All CES algorithm Unix files are stored in the CES file system, whose root directory is `/erd/statistic/project/ces`. The CES root directory name is stored in the Unix environment variable CESHOME.

Table 1 lists the principal computer program subdirectories and their purposes.

Table 1. Principal algorithm computer program subdirectories.

Directory	Description
`sas`	SAS computer programs
`macros`	SAS macros used by the SAS programs
`review`	Computer programs used for WWW review of the recommendations
`wwwssd`	Data used for WWW review of the recommendations
`library`	Custom SAS formats used for labeling plots and output

6.2.1 Algorithm Run Identification

Each algorithm run is uniquely defined by the set of wells in the run, the analytes that are evaluated at those wells, and the planning quarter. The set of wells is called a "site" and the analytes are referred to as a "www group". (The term "www group" is used in relation to the WWW review phase of an algorithm run.)

In the computer programs the site, www group, and planning quarter are referred to using abbreviations, or "codes". These codes are used in directory and file names in order to help keep the system organized and to identify the purposes and contents of directories and files. Table shows examples of some identification codes.

Table 2. Examples of algorithm run identification codes.

Site	Site code	Planning quarter	Quarter code	WWW group code
Livermore	`s200`	`1999-3`	`99-3`	(not used)
Livermore	`s200`	`1999-4`	`99-4`	(not used)
Site 300, EFA/WFA	`ewfa`	`2000-4`	`00-4`	`voc`
Site 300, EFA/WFA	`ewfa`	`2000-4`	`00-4`	`othr`

6.2.2 Algorithm Run Directories

The site, www group, and planning quarter codes are used to name the algorithm run directories.

There is a root directory for each site. For the Livermore Site, the root directory is named `$CESHOME/s200`. For Site 300, the root directories have names of the form `$CESHOME/s3.sitecode`. For example, runs for the EFA/WFA study area at Site 300 are in the directory named `$CESHOME/s3.ewfa`.

At the Livermore Site, there is only one analyte group, so the www group code is not used. Within the Livermore Site root directory there is a directory for each planning quarter. The working directories have names of the form `sfyy-q` where `yy-q` is the quarter code. See Table 2 for examples.

At Site 300, some sites have more than one www group. Within such Site 300 root directories there are directories for each www group and planning quarter. Working

directories for Site 300 have names of the form `$CESHOME/s3.sitecode/wwwcode/sfyy-q`, where `yy-q` is the quarter code, `wwwcode` is the www group code, and `sitecode` is the site code. If there is only one www group, the intermediate directory (`wwwcode`) may be omitted. See Table 2 for examples.

It is not required that these naming conventions be followed precisely. It is strongly recommended that the directory names incorporate the site, www group, and planning quarter codes.

Table 2 shows some example working directory names.

Table 2. Example of algorithm run working directory names.

Working directory name	Description
`$CESHOME/s200/sf99-3`	**Livermore Site run for VOCs, planning for Q1999-3**
`$CESHOME/s200/sf99-4`	**Livermore Site run for VOCs, planning for Q1999-4**
`$CESHOME/s3.ewfa/voc/sf00-1`	**Site 300 EWFA run for VOCs, planning for Q2000-1**
`$CESHOME/s3.ewfa/othr/sf00-1`	**Site 300 EWFA run for analytes other than VOCs, planning for Q2000-1**

6.3 Algorithm Custodian Tasks

The following is a concise summary of the steps to follow.

1. Verify that the list of wells has been reviewed and, if necessary, updated.
2. Create a new working directory (within the appropriate site and www group directory).
3. Move to the new working directory.
4. Create new subdirectories for input, output and data files.
5. Copy command and control files from the previous quarter's directory.
6. Edit the files to reflect the current quarter.
7. Review the control files and make any necessary changes.
8. Run the SAS program `getdat.sas`.
9. Run the SAS program `runalg.sas`.
10. Convert the QC chemist review file to spreadsheet format and deliver it to the QC chemist for review.
11. Receive the edited spreadsheet from the QC chemist and save it in tab delimited text format.
12. Run the SAS program `runtl.sas`.
13. Convert the project staff review file to spreadsheet format and deliver it to the project staff.
14. After the previous steps are complete (and when it is highly unlikely that step 9 will be repeated), run the SAS program `addbien.sas`.

The following sections provide details and indicate who is responsible for each step.

6.3.1 Verify that the Well List has been Reviewed (Section 6.3, step 1)

To begin an algorithm run the algorithm custodian should ask the sampling plan custodian to give the project staff a list of wells currently designated for the algorithm run. The project staff should review this list. If there are changes the project staff should inform the sampling plan custodian, who will update the list. The sampling plan custodian then informs the algorithm custodian that the list is ready.

The lists of wells for each site are maintained in the database table `locgroups`. Location group names start with the string "`alg`". Some example include `alg_ls` (Livermore Site), `alg_ewfa` (Site 300 EWFA), and `alg_834` (Site 300 B834). New location groups should be created when needed for new sites.

6.3.2 Create Working Directory (Section 6.3, steps 2, 3, 4)

The algorithm custodian performs this step.

Log in to a Unix computer in the ERD network and move to the appropriate site and www group directory (see Section 6.22).

Create the new working directory in the appropriate site and www group directory (see Section 6.22).

Move to the working directory and create four subdirectories named `raw`, `ssd`, `in`, and **`out`**. The computer programs expect to find subdirectories with these names, therefore these names *must* be used. The contents of these subdirectories are:

raw Contains SAS format files of chemical data as retrieved from the database and prepared for the algorithm run.

ssd Contains SAS format files created by the algorithm run. These include the data as used by the algorithm run, summary statistics, and the raw recommendations.

out Contains ascii (tab delimited text) format files output by SAS and used to create the spreadsheets for the reviewers.

in Contains ascii (tab delimited text) format versions of the spreadsheets after they have been reviewed.

6.3.3 Prepare Command and Control Files (Section 6.3, steps 5, 6, 7)

The algorithm custodian prepares the necessary command and control files, as described below. If necessary, log in to a Unix computer in the ERD network and move to the working directory.

The easiest way to prepare the files is to copy them from the previous planning quarter's directory for the same site and www group, and edit them to reflect the new planning quarter.

There are two categories of files necessary to perform an algorithm run: command files and control files. In addition, there should be an informational file named README.

Command files contain the computer programs that perform the algorithm run. The required command files are:

```
autoexec.sas   runtl.sas
getdat.sas  addbien.sas
runalg.sas
```

The name `autoexec.sas` is required by the SAS software system and must not be changed. The other names should not be changed.

Control files contain information that controls various aspects of how the run is performed. The required control files are

`control.sas`
`cutoffs`
`mcls`
`guard`

The name `control.sas` is required and must not be changed.

The other control files have names that depend on the site. At the Livermore Site, they are named `cutoffs.s200`, `mcls.s200`, and `guard.s200`. At other sites, different file name suffixes may be used, but the first part (`cutoffs`, `mcls`, and `guard`) should not be changed. If necessary, edit `control.sas` to reflect the names that are used.

Inspect the README file from the previous quarter for any notes that may have relevance for the planning quarter. Add notes as needed. Update it to refer to the planning quarter.

In the SAS command files (`autoexec.sas`, `getdat.sas`, `runalg.sas`, `runtl.sas`, and `addbien.sas`) update all references to the planning quarter. Use the format `yyyy-q` for the quarter (e.g., `1999-1`).

In `control.sas`, update the values of the variables `basedate`, `cutdate`, `shistlo`, and `shisthi`. Normally, `basedate` is one month before the beginning of the planning quarter and `cutdate` is the beginning of the planning quarter. Normally, `shisthi` is the year of the previous quarter and `shistlo` is 8 years before. For example, if planning for Q1999-1, then `shistlo=1991` and `shisthi=1998`, but when planning for Q1999-2 then `shistlo=1992` and `shisthi=1999`.

Review the names of the input and output files defined in `control.sas`. Most of the file names are constructed from the site, www group, and quarter codes, and normally should not be changed.

If instructed to do so by the algorithm expert or statistician, edit additional variables in `control.sas` or the values in the `cutoffs`, `mcls`, and `guard` files, and make a note in the README file.

Normally, data is retrieved from the most recent Gemini database. `getdat.sas` is configured to automatically determine which Gemini database is the most recent. If necessary, `getdat.sas` may be edited to use a particular Gemini database, or the production database, EPDDATA.

If this is a run for a new site or www group then `getdat.sas` will require editing in order to retrieve the correct data from the database. If necessary, consult with the sampling plan custodian or member of the data management staff.

All of the command files should have comments near the beginning of the file that include the filename, the date the file was created, and any other comments that may be useful. It is good practice update these.

6.3.4 Retrieve Data from the Database (Section 6.3, step 8)

The algorithm custodian performs this step. If necessary, log in to a Unix computer in the ERD network and move to the working directory.

Execute the SAS program that retrieves data from the database by typing the Unix command

```
sas getdat
```

This step retrieves data for selected wells and chemicals from the most recent `gemini` database. The SAS program determines which `gemini` database is the most recent.

Review the file **getdat.log** for error and warning messages.

The files `getdat.lst` should be inspected for indications of problems. If there are wells that are specified in the database as part of the algorithm run, but for which there is no data, they will be listed in `getdat.lst`.

6.3.5 Calculate Raw Recommendations (Section 6.3, step 9)

The algorithm custodian performs this step. If necessary, log in to a Unix computer in the ERD network and move to the working directory.

Execute the SAS program that calculates raw recommendations by typing the Unix command

```
sas runalg
```

This step fits statistical models to the data, compares the results to criteria in the cutoffs, guard, and mcls files, and determines recommended sampling frequencies.

Review the file `runalg.log` for error and warning messages.

Review output files for indications of problems. In particular, inspect the files `out/abbrev.lst`, `out/uniterr.lst`, and `out/coderr.lst`. Samples with inconsistent units, if any, will be listed in `uniterr.lst`. This file may be empty, indicating that all units are correct. The files `abbrev.lst` and `coderr.lst` will indicate discrepancies, if any, between analyte names used in the input data and analyte names used in the `cutoffs` file. This is especially important to review when preparing files for a new www group.

6.3.6 Prepare for Review by QC Chemist (Section 6.3, step 10)

The algorithm custodian performs this step. If necessary, log in to a Unix computer in the ERD network and move to the working directory.

Convert the QC chemist review files to Excel format and deliver them to the QC chemist.

The output file names are found in control.sas, as the values of the macro variables "`ssout`", "`chmout`", and "`ndchmout`". For example, the "`ssout`" spreadsheet might be named `LS2000-4.a`, as a result of the statement "`%let ssout=out/LS&qtr..a;`" in `control.sas`.

6.3.7 Review by QC Chemist (Section 6.3, step 10)

The QC chemist performs this step.

The "`ssout`" spreadsheet contains a column whose heading is "`Rev`" (an abbreviation of "Review"). The QC chemist enters his or her review decisions in this column. Entries are made only in cells with a "~" in them. When the QC chemist does not change a recommendation, the "~" is left unchanged.

The QC chemist may change the spreadsheet name.

The other spreadsheets ("chmout" and "ndchmout") are for information only; they contain results from the algorithm run that the QC chemist may wish to consult during review.

The QC chemist returns the edited spreadsheet to the algorithm custodian.

6.3.8 Receive Results of QC Chemist Review (Section 6.3, step 11)

The algorithm custodian performs this step. If necessary, log in to a Unix computer in the ERD network and move to the working directory.

Make a tab-delimited ascii copy of the edited spreadsheet and put it in the subdirectory 'in'. The file name should be the same as the spreadsheet name, but with the suffix '.a' appended.

Set the variable revin in runtl.sas to the name of the converted spreadsheet file (the variable revin may optionally be set in control.sas instead).

6.3.9 Prepare for Review by Project Staff (Section 6.3, steps 12, 13)

The algorithm custodian performs this step. If necessary, log in to a Unix computer in the ERD network and move to the working directory.

Verify that the input file name in is correct (see Section 6.3.8).

Execute the SAS program that prepares for project staff review by typing the Unix command

```
sas runtl
```

Convert the project staff review file to Excel format and deliver it to the project staff for review. The output file name is found in runtl.sas as the value of the SAS macro variable "tlss" (the variable tlss may optionally be set in control.sas instead).

This step also prepares SAS datasets for use by the graphical review program that is accessed via the WWW. The WWW review page is within the password protected "ERD Only" internal web pages.

When project staff review is complete, the project staff deliver the reviewed recommendations to the Sampling Coordinator and the Sampling Plan Custodian, and inform the Algorithm Custodian.

6.3.10 Append Raw Recommendations (Section 6.3, step 14)

The algorithm custodian performs this step. If necessary, log in to a Unix computer in the ERD network and move to the working directory.

After all of the above steps are complete, and it is highly unlikely that program runalg.sas will be rerun (See Section 6.3.0), the raw recommendations must be uploaded to the database. This makes them available for the next algorithm run. In order to perform this step the algorithm custodian must have insert privileges on the database table ces_raw_rec.

Type the Unix command

```
sas addbien
```

6.3.11 Data Storage

The computer program files, log files, output files, and input files are kept online for up to one year, after which they are kept on backup tapes.

7. QA RECORDS

The final sampling plan, created after the project staff have completed their review (see ERD SOP-2.11), is the QA Record for this procedure.

8. ATTACHMENTS

Not applicable.